THE USE of MAJOR RISK FACTORS for COMPUTER-BASED DISTINCTION of DIABETIC PATIENTS with ISCHEMIC STROKE and WITHOUT STROKE

Sibel Oge Merey[1], Fikret Gurgen[2], Nurgul Gurgen[2]
[1]Electronics and Electrical Eng. Dept., Technical Univ of Istanbul, Maslak-Istanbul/TURKEY
[2]Dept of Computer Eng., Bo□aziçi University, TR-80815 Bebek-Istanbul/TURKEY
[3]Neurology Division, Lütfiye Nuri Burat Hospital, Sultanciftli□i Istanbul/TURKEY
E-mail: gurgen@boun.edu.tr

Abstract- **This study proposes a computer-based decision support system to investigate the distinctive factors of diabetes mellitus (DM) with ischemic (non-embolic type) stroke and without stroke. Database consists of a total of 16 features that are collected from 44 diabetic patients. Features include age, gender, duration of diabetes, cholesterol, higher density lipoprotein (HDL), triglicerit levels, neuropathy, nephropathy, retinopathy, peripheral vascular disease (PVD), myocard infarction (MI) rate, glucose levels, taking medicine, blood pressure. Metric and non-metric features are distinguished. First, the statistics, mean and covariance, of data are estimated and the correlated components are observed. Second, principal component analysis (PCA) is used for major components. Finally, decision making approaches, k-nearest neighbor (k-NN) and MLP, are employed for classification of all the components and major components case. Macrovascular changes emerged as principal distinctive factors of ischemic-stroke in DM. Microvascular changes were generally ineffective discriminators. Recommendations were made according to the rules of evidence-based medicine. Briefly, this case study supports theories of stroke in DM and also concludes that the use of intelligent data analysis improves personalized prevention.**

Key words: **Intelligent Data Analysis (IDA), decision support system, diabetes mellitus (DM), non-embolic (ischemic) stroke, principal component analysis (PCA), decision making, k-nearest neighbor (k-NN), multilayer perceptron (MLP)**

1 INTRODUCTION

Stroke has been an important health issue and expressing and interpreting risk factors provides vital information [1-13]. This study discusses a computational method for highlighting the major risk factors of diabetic patients with non-embolic stroke and without stroke by performing dependency analysis and decision making. For this purpose, the follow up data of 22 diabetic patients with ischemic stroke (non-embolic type) and 22 diabetic patients without stroke were collected during a few years. Average population age was 66.2 ± 9.9 for the stroke group and 61 ± 6.1 for the control group. It is known that diabetes mellitus (DM) is diagnosed by fasting glucose level is higher than 140 mg/dl and random glucose level is higher than 200 mg/dl in repeated measurements. The study population of 44 patients were chosen with these glucose levels. Then, a set of tests are applied to construct the parameters of each feature vector. The tests include age, gender, duration of diabetes, cholesterol, higher density lipoprotein (HDL), trigliserit levels, neuropathy, nephropathy, retinopathy, peripheral vascular disease, myocard infarction rate, fasting and random glucose levels (FGL and RGL), taking medicine, sistolic and diastolic blood pressure. As clearly observed, feature vectors contain metric and nonmetric components. For example, a blood cholesterol level test is a metric component that all mathematical operations can be performed with the highest level of precision. On the other hand, retinopathy is a nonmetric component that provides a nominal scale to label or to identify to retina conditions.

This article proposes a decision support system and provides the physician with an objective way of knowing main risks of stroke in a diabet patient concerning the pathologies for which preventive measures exist. The purpose is to use existing knowledge for developing prevention strategies based on evidence-based medicine [13] that fits the patient and complies with current scientific data. Thus, this system transforms data into information that is used for decision making together with expert knowledge.

2 DIABETIC PATIENTS WITH NON-EMBOLIC STROKE AND WITHOUT STROKE

Monitoring DM patients with various diagnostic tests and searching evidences of disease are generally routine and critical also in stroke [4-5]. Measuring major risk factors of ischemic stroke in DM can also provide arguments to the attending physician to initiate preventive measures adapted to patient's case. Typical microvascular complications are neuropathy, nephropathy, retinopathy and macrovascular ones are coroner artery diseases (CAD) and peripheral vascular diseases (PVD) [6-7]. As known, microvascular and macrovascular complications may occur during DM. The macrovascular complications such as cholesterol, HDL, trigliserit levels, FGL and RGL, sistolic and diastolic blood pressure are considered the risk factors of nonembolic-ischemic stroke on DM subjects [5-12]. Various studies have addressed the relationship between DM and stroke, and the probable risk factors [4-12]. Even though ischemic cerebravascular is generally caused by macrovascular changes [7-9], the cerebral ischemy is claimed by the small artery occlusions in DM [6]. In DM, age, hypertension, MI, PVD are the known risks of stroke [4-5]. Also, fewer studies are available to indicate the relationship between neuropathy, nephropathy, retinopathy and stroke.

Report Documentation Page

Report Date 25 Oct 2001	**Report Type** N/A	**Dates Covered (from... to)** -

Title and Subtitle The Use of Major Risk Factors for Computer-Based Distinction of Diabetic Patients with Ischemic Stroke and Without Stroke	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) Electronics and Electrical Eng. Department Technical Univ of Istanbul Maslak-Istanbul/Turkey	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)
Distribution/Availability Statement Approved for public release, distribution unlimited	
Supplementary Notes Papers from 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom., The original document contains color images.	
Abstract	
Subject Terms	
Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU
Number of Pages 5	

Means, standard deviations and correlations are estimated from the observed population. It is known that small sample sizes reduce the statistical power of the study and generally result in wide confidence interval of estimated parameters. Thus, identical studies on larger samples may identify more important differences that have gone undetected here. Also, the deviation from the normality may cause errors in parameter estimations. As a compensation to hardly-collected, small size of study population problem, we first observe these statistics by normal distribution assumption then, attempt to employ nonparametric classification techniques such as k-NN and MLP with jackknife method. Even though there exist ways of checking the fitness: for example, how well the normality of distribution for each component can be evaluated by various tests such as Kolmogorov-Smirnov test. But our aim is to build a decision support mechanism.

Regardless of the limitations imposed by technical artifacts and sample size constraints, our results describe several important findings emerged from this preliminary investigation of the relationship of diabet patients between non-embolic stroke and without stroke.

Mean and standard deviation of metric components: cholesterol, HDL, trigliserit levels, FGL and RGL, sistolic and diastolic blood pressure were found (Table 1). Then, correlation estimate denoted pairwise relationships of the components. Nonmetric components were counted as the percentage of existence or nonexistence cases (Table 1).

3 INVESTIGATING PRINCIPAL FACTORS OF DATA

By PCA analysis [14-18] of all components, we find the first macrovascular components of Table 1: cholestrol, trigliserit, glucose levels, blood pressures as the principal components. In decision making point of view, we observe that they are the most informative components in the study population. In fact, this confirms the reports of major risk factors in the literature [4-13]. Among the other components, the choice of age and gender are designed during data collection and have less information. Microvascular changes, neuropathy, nephropathy, retinopathy, are also assessed by only two distinctive levels (existence-nonexistence) to show the subgroups of DM subjects. Thus, they have a limited information that can be used for decision. PCA is found to be valuable for summarizing the trends of features of high risk subjects with DM.

4 DECISION MAKING BY PRINCIPAL FACTORS

It is known that there is no standard approach to establish the optimum decision criterion for a generic problem. Nonparametric classification approaches [15-17] overcome the difficulty of accurate estimation of underlying distributions from small population size. They are used without assumption about form of density function and their decision procedures bypass probability estimation and go directly to decision functions. Due to small population, we here employ nonparametric techniques instead of parametric ones such as K-Means, vector quantization (VQ) and gaussian mixture model (GMM) which, otherwise, would be very useful with the large population. Furthermore, fuzzy modelling methods enhance the decision ability [19].

Among various nonparametric techniques, we choose the local k-nearest neigbor (k-NN) and the global multilayer perceptron (MLP) approach. The concept of locality and globality is related to the location of information that is extracted for class decision. k-NN rule employs local information, in contrast, MLP extracts global information.

In k-NN rule, the class decision of unknown sample is based on the majority of the nearest k samples. In other way, a local voting process decides the class of unknown with the highest vote. This is, in fact, an approximation of Bayes Decision Theory in a local environment [16]. This process can further be extended to weighted voting and gaussian weighted voting (Parzen window) [16-18]. To recognize pattern v, k minimum distance samples are computed among all the samples. The distance can be computed by various norms: minkowski-norm-based distances, e.g. euclidian as second order, covariance-based distance or mahalonobis, entropy-based distance or kullback-leibler distances. The simplest and the most suitable for small sample size is euclidian metric which is defined as

$$dj = |v-v_j|^2 = (v- v_j)^T(v- v_j) \qquad (1)$$

where dj defines the distance between patterns v and $v_{j.}$ k-NN rule classifies v by assigning it a label that is the most frequently represented among the k nearest samples:

$$ki = \max \{k_1,\ldots,k_L\} \qquad \underset{k_1+\ldots+k_L = k}{x \,\square\, w_i} \qquad (2)$$

Here, k_i is the number of neighbors belong to w_i (i=1,….L) class among the k nearest neighbor.

The global MLP [15-18] is a parallel, feedforward structure that consists of input, output and hidden layers. Each layer has sigmoidal units interconnected through weight connections. The MLP is trained with supervised back propagation (BP) algorithm to efficiently compute partial derivatives of an approximating function F(w;x). The network has an adjustable weight vector w that is computed with respect to all training data for a given value of input vector x and output vector y. The weights are adjusted to fit a set of surfaces to the input space. The surfaces are constructed by sigmoids by the best linear regression concerning the cluster membership. The mean square (MSE) (5) error function, which is the difference between the network's output and the supervisor output, is minimized to find the cluster membership:

$$MSE = \sum_{q} (y_q - F(w; x_q))^2 \qquad (3)$$

In k-NN rule, an arbitray discriminant function is constructed for class decision. The nearest neighbor contributes the half of the classification information. A MLP can generate any nonlinear discriminant function of input by incorporating multiple constraints. Each sigmoidal unit contributes to global discriminant function by a linear constraint.

5 TESTS AND RESULTS

The proposed system is shown in Fig. 1. Figure 2 shows the distribution of RGL-FGL-cholestrol features. The study population contains 44 diabetic patients' collected by neurology specialists. The data has a total of 16 metric and non-metric components. The computer-based system uses two stages: first, PCA method reduces the features to 7 dimensions, then, the best discriminators for class membership are searched by nonparametric classifiers, k-NN and MLP. Multi-parameter classifiers were found to be significantly improve upon the classification performance of single parameter designs. Instead of single parameter based conclusions, we employed the decision produced by major risk factors. This makes the study have stronger arguments on the distintive factors. The small sample population size has an acknowledgable potential effect on the statistical power of the study but we use classification techniques like k-NN and MLP to overcome this drawback.

The results are shown in Table 3 and Table 4. An average classification score of 68.18% is obtained for k-NN. The performance of MLP is tested using hold-one-out method (jackknife) as 100%. Both classifiers identify the macrovascular changes as the best discriminators as of this case study on stroke.

The figure of component clusters clearly provides convincing visual information to the physician and patients as the end-users. This easy-to-use graphic offers users information about the limits of the major risk factors.

6 DISCUSSIONS AND CONCLUSION

In summary, this paper presents a computer-based decision support system for physicians to improve prevention measures for ischemic stroke in DM subjects. PCA helps to identify the basic trends that distinguish two cases: ischemic stroke and no stroke in DM. In this study, major distinctive factors are found to be cholesterol, trigliserit level, fasting and random glucose levels (FGL and RGL), sistolic and diastolic blood pressure levels. Physicians furthermore can suggest diatery supplies, drugs, exercises, etc. for the prevention of ischemic stroke for diabetic patients with these findings. Visual graphics of the components also support the discussion for clues of prevention.

The aim of this study is to support doctors with intelligent data analysis techniques. The system may be thought as an indirect assistance tool for physicians. We also plan to improve the information of diagnosis with a certain risk assessment factor. A grading for the risk of ischemic stroke for diabetic patients will support the prevention measures according to evidence-based medicine.

The major concern of the study is that the investigators could not have a bigger size of study population. This raises concerns about the meaningfulness of results but the authors believe that the methods used here clearly useful to overcome this handicap. Also the results of the study are supported by the risk factor findings of the other stroke and diabet studies in the literature [4-13]. When there is a large database, this will provide more convincing statistical power to the study . In thi,s case, the general statistics will confirm the results and parametric decision making techniques such as expectation-maximization (EM) algorithm would be utilized.

The other limitation is the inability to directly account for certain microvascular factors in the decision process due to imprecision of their records: for example, retinopathy has only two identifying labels as exist or non exist. In fact, there are typically four cases that can be distinguished by the physician but the limited data size prevents us to include finer details. This may cause an error. The lack of precision, in general, reduces the effect of the feature in decision making.

Among the various other tests, we also believe that the doppler ultrasound measurements may present more detailed information on this case study. The areas of arteries and veins can be monitored for deciding the stages of the disease, and the condition of occlusions can be important for the preventive follow up.

The other weak point of this investigation may be the quality issues of the recording such as the use of nonstandardized procedures, the use of potentially low grade equipment or the measurement quality. However, as best supported by a group of physicians' observations, this was not the case.

As a conclusion, this case study points at a friutful line of enquiry in intelligent medical data analysis and the content of the work can further be extended to the other areas of stroke diagnosis and prevention.

7 REFERENCES

[1] Brause, Rudiger, "Medical Data Analysis," Springer-Verlag, 2000.

[2] Hand D. J., J. N. Kok, M. R. Bertholt, "Advances in Intelligent Data Analysis," Springer-Verlag, 1999.

[3] Gurgen F., “Neural-Network-Based Decision Making in Diagnostic Applications”, IEEE Engineering in Medicine and Biology, pp 89-93, July/August, 1999.

[4] Barnett H. J. M., Mohr J. P., Stein B. M., Yatsu F. M., “Stroke: Pathophysiology, Diagnosis and Management,” Secon Edition, Churchill Livingstone Inc., 1992.

[5] Bogousslavssky J., Caplan L., “Stroke Syndromes,” Cambridge University Press, 1995.

[6] Nurgül Ayd□n, Haluk Esgin, Aynur Y□lmaz, Filiz Gözeten, Ufuk Utku, Diabetes Mellitus'lu Non-Embolik Stroklu Olgularda Retinopati ve Di□er Risk Faktörleri, *Türk Beyin Damar Hastal□klar□ Derne□i 3. Sempozyumu*, 1999.

[7] David S. H., Bell, MB. "Stroke in the diabetic patient. Diabetes Care" 17, pp. 213-219, 1994.

[8] Jose B, MD, FACP, Betsy B. Love, MD. "Diabetes and stroke. Medical Clinics of North America" 77, number 1, pp. 95-109, 1993.

[9] Diana B. P., Hilarey B., "Retinopathy as a Risk Factor for Nonembolic Stroke in Diabetic Subjects. Stroke," Vol. 26, pp. 593-596, 1995.

[10] Jose B., Betsy B. L., "Diabetes and Stroke," Medical Clinics of North America, Vol. 77, No 1, pp. 95-109, 1993.

[11] Abbott R. D., Donahue R. P., " Diabetes and Risk of Stroke," JAMA, Vol 257, pp. 949-952, 1987.

[12] Mortel K. F., Meyer J. S., Sims P. A., McClintic K., "Diabetes Mellitus as a Risk Factor for Stroke," Southern Medical Journal, Vol. 83, pp. 904-911, No. 8.

[13] Sackett D. L., Rosenberg W. M., Gray J. A., Haynes R. B., and Richardson W.S., "Evidence-based Medicine: What it is and what it isn't," Mr. Med. J., Vol 312, pp. 71-72, 1996.

[14] S. Sharma, "Applied Multivariate Techniques," John Wiley, 1996.

[15] A. C. Rencher, "Methods of Multivariate Analysis," John Wiley, 1995.

[16] Richard O. Duda, Peter E. Hart, "Pattern Classification and Scene Analysis*", 1972.*

[17] Keinosuke Fukunaga, "Introduction to Statiscal Pattern Recognition," 1990.

[18] Jürgen Schürmann, "Pattern Classification, A Unified View of Statistical and Neural Approaches," 1996.

[19] T.Takagi and M. Sugeno: "Fuzzy identification of systems and its applications to modelling and control", *IEEE Trans.* Systems, Man and Cybernetics, 15(1), 116-132, 1985.

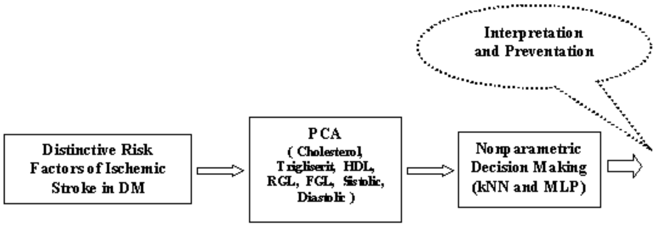
Interpretation and Prevention
Distinctive Risk Factors of Ischemic Stroke in DM
PCA (Cholesterol, Trigliserit, HDL, RGL, FGL, Sistolic, Diastolic)
Nonparametric Decision Making (kNN and MLP)

Figure 1: Distinctive risk factors of stroke in DM and without stroke

	Standard deviation and mean value			
	Ischemic stroke		No stroke	
	St. Dev.	Mean	St. Dev.	Mean
Age	9.8992	66.2273	6.1101	61
Cholesterol	54.5235	203.3636	50.9238	217.7727
Trigliserit	97.7951	150.5	89.8721	155.9545
HDL	6.4842	40.9545	6.8307	44.0909
RGL	86.168	206.5909	53.0226	196.1818
FGL	75.3457	187.2727	62.5272	180.3182
Sistolic	26.2851	143.6364	14.6015	143.1818
Diastolic	12.9267	83.6364	11.0195	85
DM (month)	91.0637	97.8636	87.6972	186.3182

Table 1 Statistics of metric components of DM patients with ischemic stroke without stroke

	first class	second class

Gender
Female
Male
	6 (%27.3)
	16 (%72.7)
		8(%36.4)
		14(%63.6)

Medicine
Used
Not used
	11 (%50)
	11 (%50)
		17(%77.3)
		5(%22.7)

Neuropathy
Exist
Nonexist
	15 (%68.2)
	7 (%31.8)
		19(%86.3)
		3(%13.7)

Nephropathy
Exist
Nonexist
	7 (%31.8)
	15 (%68.2)
		9(%40.9)
		13(%59.1)

PVD
Exist
Nonexist
	6 (%27.3)
	16 (%72.7)
		7(%31.8)
		15(%68.2)

Retinopathy
Exist
Nonexist
	4 (%18.2)
	18 (%81.8)
		2(%9.1)
		20(%90.9)

MI
Exist
Nonexist
	4 (%18.2)
	18 (%81.8)
		1(%4.5)
		21(%95.5)

Table 2 Statistics of non metric components of DM patients with ischemic stroke without stroke

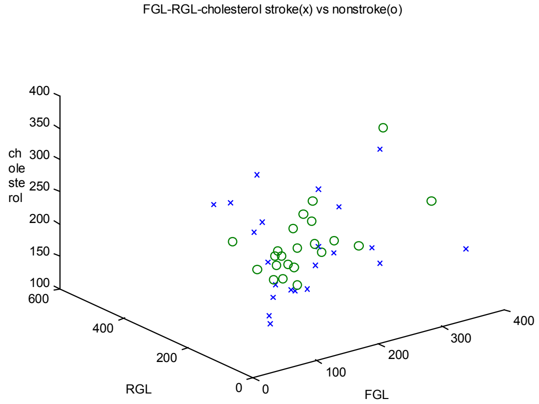

Fig. 2 Apperance of RGL-FGL-cholestrol data

	All Components	Principle Components
	% for total	
k=1	52.2727	52.2727
k=3	65.9091	65.9091
k=5	68.1818	68.1818
k=7	63.6364	63.6364
k=10	68.1818	68.1818
k=20	63.6364	63.6364
k=40	29.5455	29.5455

Table 3 Results of k-NN method

Epsilon=0.1 alpha=0.8 iteration=10000			
success rate (%) for the			
h	**first class**	**second class**	**whole**
2	50	100	75
4	86.36	86.36	86.36
6	100	100	100

Table 4 Results of MLP method